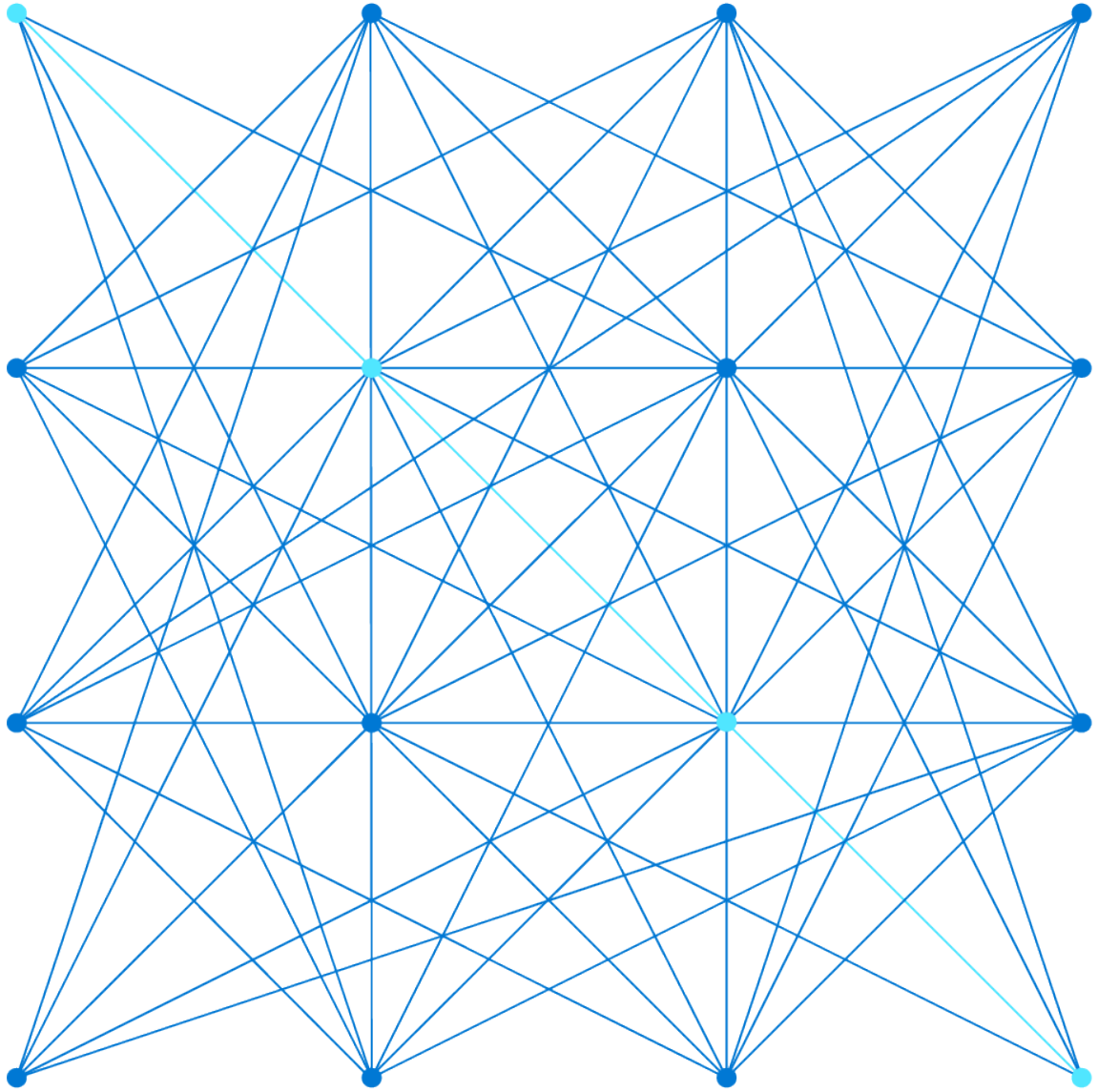
Microsoft Azure

4: Explore fundamentals of data analytics



© Copyright Microsoft Corporation. All rights reserved.

Agenda

Large-scale data warehousing

Streaming and real-time analytics

Data visualization

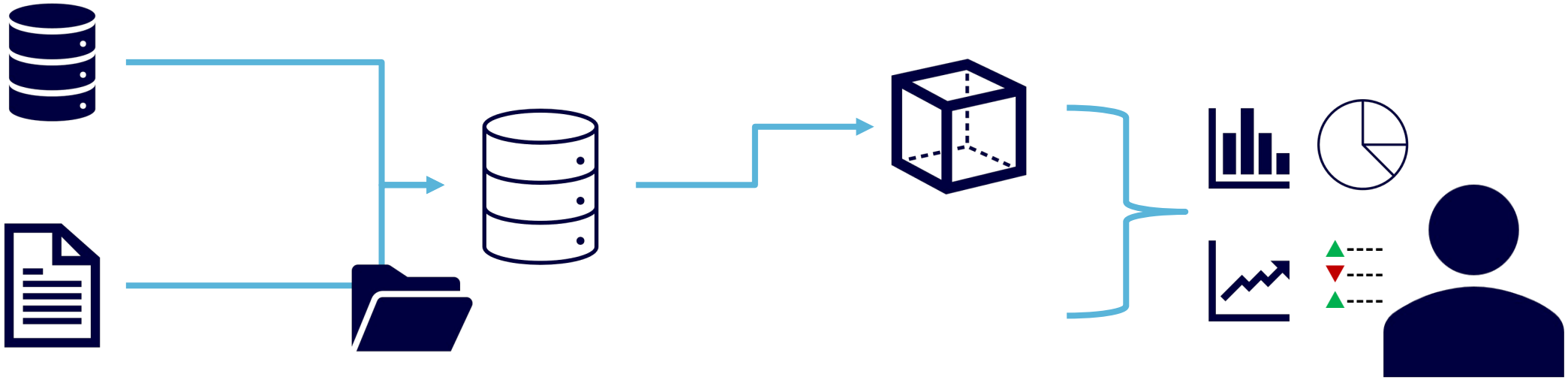
© Copyright Microsoft Corporation. All rights reserved.

1: Large-scale data warehousing

© Copyright Microsoft Corporation. All rights reserved.

What is large-scale data warehousing?

Data ingestion and processing	Analytical data store	Analytical data model	Data visualization



- *Extract, Transform, and Load* (ETL) or *Extract, Load, and Transform* (ELT) orchestration
- Distributed processing to cleanse and restructure data at scale
- Batch and real-time data processing

- Denormalized relational data storage in a *data warehouse*
- Semi-structured file storage in a *data lake*

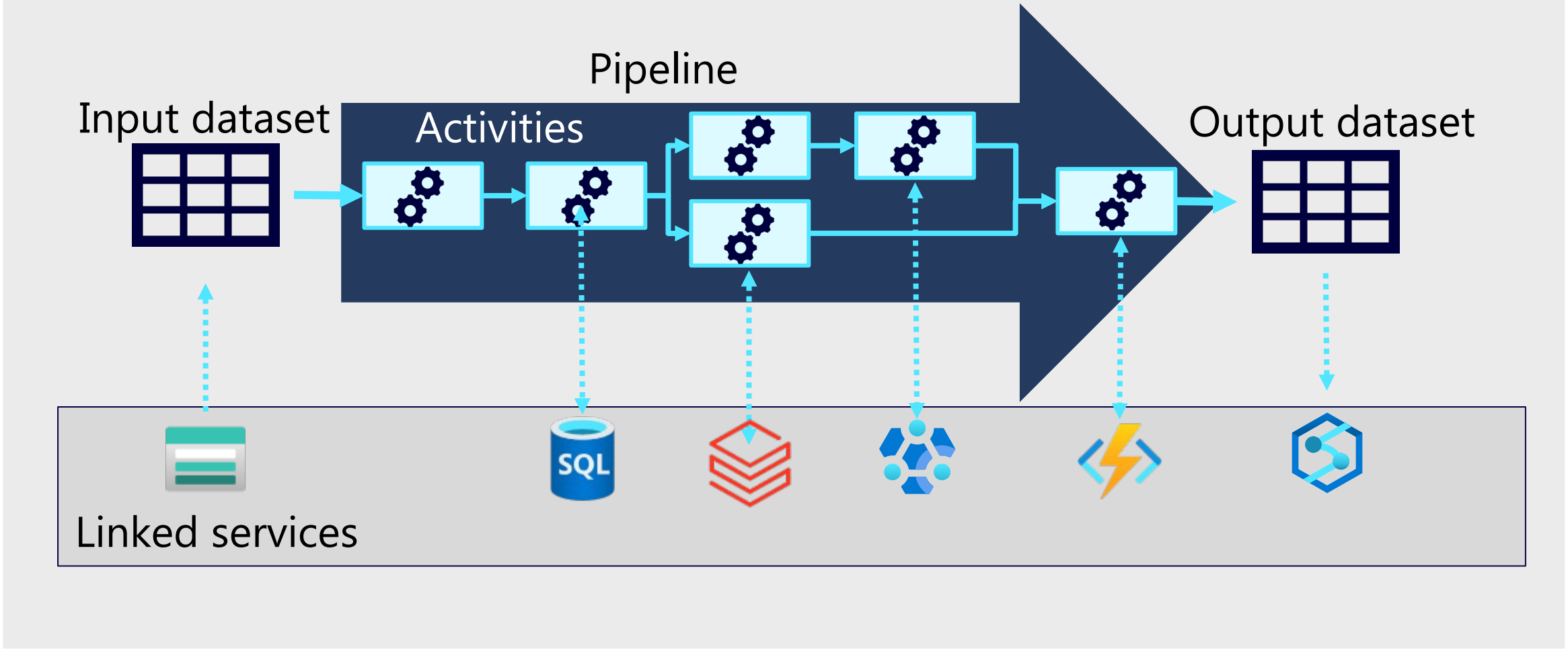
- Semantic models for analytical entities
- Often in the form of aggregated *cubes* that summarize numeric values across one or more *dimensions*

- Reports
- Charts
- Dashboards

© Copyright Microsoft Corporation. All rights reserved.

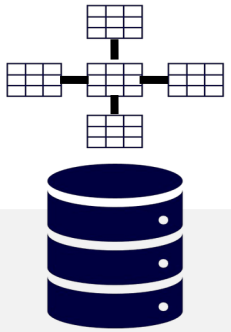
Data ingestion and processing pipelines

Create pipelines in **Azure Data Factory** or **Azure Synapse Analytics**



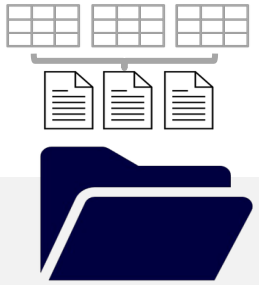
© Copyright Microsoft Corporation. All rights reserved.

Analytical data stores



Data Warehouse

- Large-scale relational database store and query engine
- Data is *denormalized* for query optimization
  - Typically as a *star* or *snowflake* schema of numeric *facts* that can be aggregated by *dimensions*



Data Lake

- Data files are stored in a distributed file system
- Tabular storage layers can be used to abstract files and provide a relational interface.
  - Use *PolyBase* external tables or create a *lake database* in Azure Synapse Analytics
  - Use database tables and SQL endpoints in Azure Databricks
  - Use Spark *Delta Lake* to add relational storage semantics and create a *data lakehouse* in Azure Synapse Analytics, Azure Databricks, and Azure HDInsight

© Copyright Microsoft Corporation. All rights reserved.

Choose an analytical data store service

Azure Synapse Analytics

- Unified solution for relational data warehouse and data lake analytics
- Scalable processing and querying through multiple analytics runtimes
  - Synapse SQL
  - Apache Spark
  - Synapse Data Explorer
- Interactive experience in Azure Synapse Studio
- Built-in pipeline integration for data ingestion and processing

Use for a single, unified large-scale analytical solution on Azure

Azure Databricks

- Azure-based implementation of Databricks cloud analytics platform
- Scalable Spark and SQL querying for data lake analytics
- Interactive experience in Azure Databricks workspace
- Use Azure Data Factory to implement data ingestion and processing pipelines

Use to leverage Databricks skills and for cloud portability

Azure HDInsight

- Azure-based implementation of common Apache "big data" frameworks built on a data lake
  - Hadoop - Query data lake files using Hive tables
  - Spark – Use Spark APIs to query data, and abstract underlying file storage as tables
  - Kafka – Real-time event processing
  - Storm – Stream processing
  - HBase – NoSQL data store

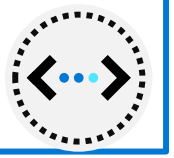
Use when you need to support multiple open-source platforms

© Copyright Microsoft Corporation. All rights reserved.

Lab: Explore Azure Synapse Analytics

In this lab, you will provision an Azure Synapse Analytics workspace, and use it to ingest and process data

1. Start the virtual machine for this lab
   or go to the exercise page at https://aka.ms/dp900-synapse-lab
2. Follow the instructions to complete the exercise on Microsoft Learn
   Use the Azure subscription provided for this lab



© Copyright Microsoft Corporation. All rights reserved.

1: Knowledge check

Which Azure services can you use to create a pipeline for data ingestion and processing?

- ☐ Azure SQL Database and Azure Cosmos DB
- ☑ Azure Synapse Analytics and Azure Data Factory
- ☐ Azure HDInsight and Azure Databricks

What must you define to implement a pipeline that reads data from Azure Blob Storage?

- ☑ A linked service for your Azure Blob Storage account
- ☐ A dedicated SQL pool in your Azure Synapse Analytics workspace
- ☐ An Azure HDInsight cluster in your subscription

Which open-source distributed processing engine does Azure Synapse Analytics include?

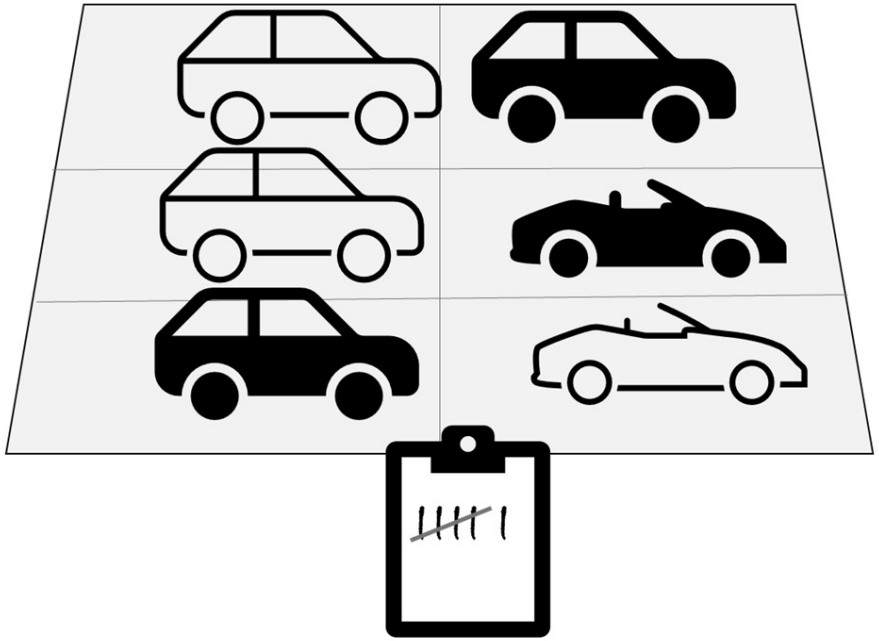
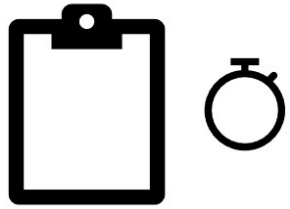
- ☐ Apache Hadoop
- ☑ Apache Spark
- ☐ Apache Storm

© Copyright Microsoft Corporation. All rights reserved.

2: Streaming and real-time analytics

© Copyright Microsoft Corporation. All rights reserved.

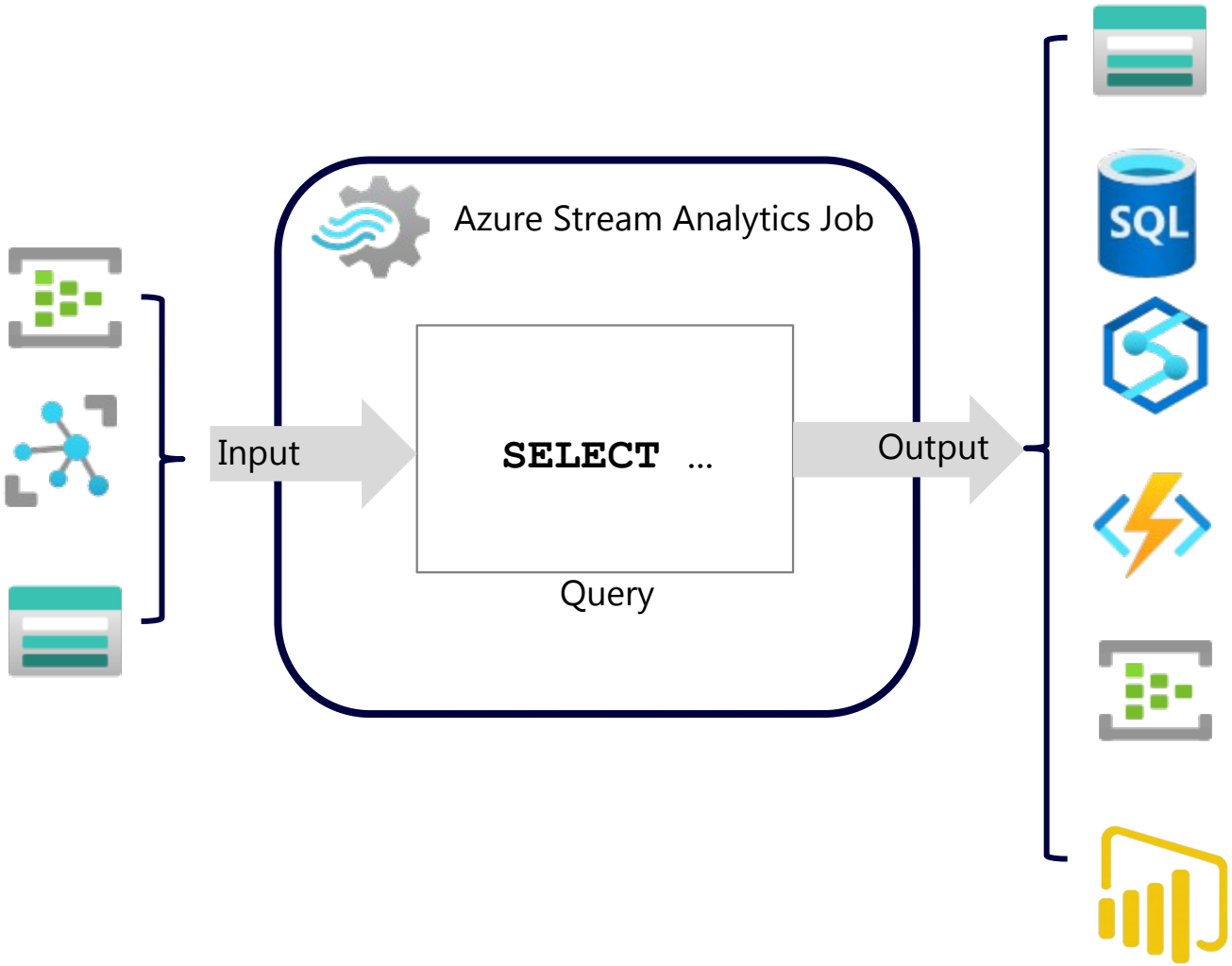
Batch vs stream processing

Batch processing	Stream processing
	
Data is collected and processed at regular intervals	Data is processed in (near) real-time as it arrives

© Copyright Microsoft Corporation. All rights reserved.

Real-time data processing with Azure Stream Analytics

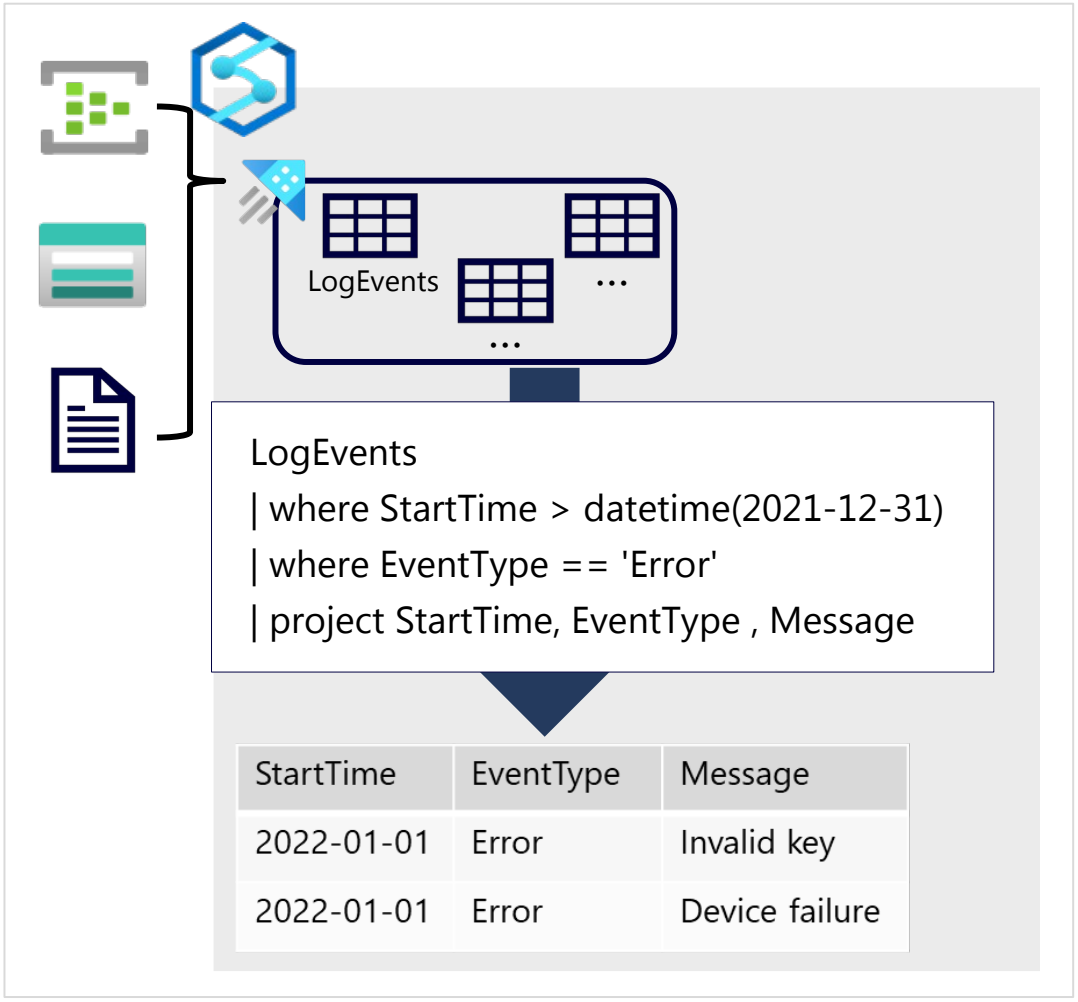
- Create an individual Azure Stream Analytics *job* or an Azure Stream Analytics *cluster*
- Ingest data from an *input*, such as:
  - Azure Event Hubs
  - Azure IoT Hub
  - Azure Blob Storage
  - …
- Process data with a perpetual *query*
- Send results to an *output*, such as:
  - Azure Blob Storage
  - Azure SQL Database
  - Azure Synapse Analytics
  - Azure Function
  - Azure Event Hubs
  - Power BI
  - …



© Copyright Microsoft Corporation. All rights reserved.

Real-time log and telemetry analysis with Azure Data Explorer

- High throughput, scalable service for batch and streaming data
  - **Azure Data Explorer** dedicated service
  - **Azure Synapse Data Explorer** runtime in Azure Synapse Analytics
- Data is ingested from streaming and batch sources into tables in a database
- Tables can be queried using *Kusto Query Language* (KQL):
  - Intuitive syntax for read-only queries
  - Optimized for raw telemetry and time-series data

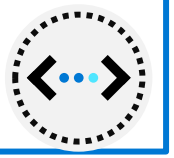


© Copyright Microsoft Corporation. All rights reserved.

Lab: Analyze streaming data

In this lab, you will use Azure Stream Analytics to process a real-time data stream

1. Start the virtual machine for this lab
   or go to the exercise page at https://aka.ms/dp900-stream-lab
2. Follow the instructions to complete the exercise on Microsoft Learn
   Use the Azure subscription provided for this lab and a cloud shell in the Azure portal



© Copyright Microsoft Corporation. All rights reserved.

2: Knowledge check

Which definition of *stream processing* is correct?

- [x] Data is processed continually as new data records arrives
- [ ] Data is collected in a temporary store, and all records are processed together as a batch
- [ ] Data is incomplete and cannot be analyzed

Which service would you use to continually capture data from an IoT Hub, aggregate it over temporal periods, and store results in Azure SQL Database?

- [ ] Azure Cosmos DB
- [x] Azure Stream Analytics
- [ ] Azure Storage

Which language would you use to query real-time log data in Azure Synapse Data Explorer?

- [ ] SQL
- [ ] Python
- [x] KQL

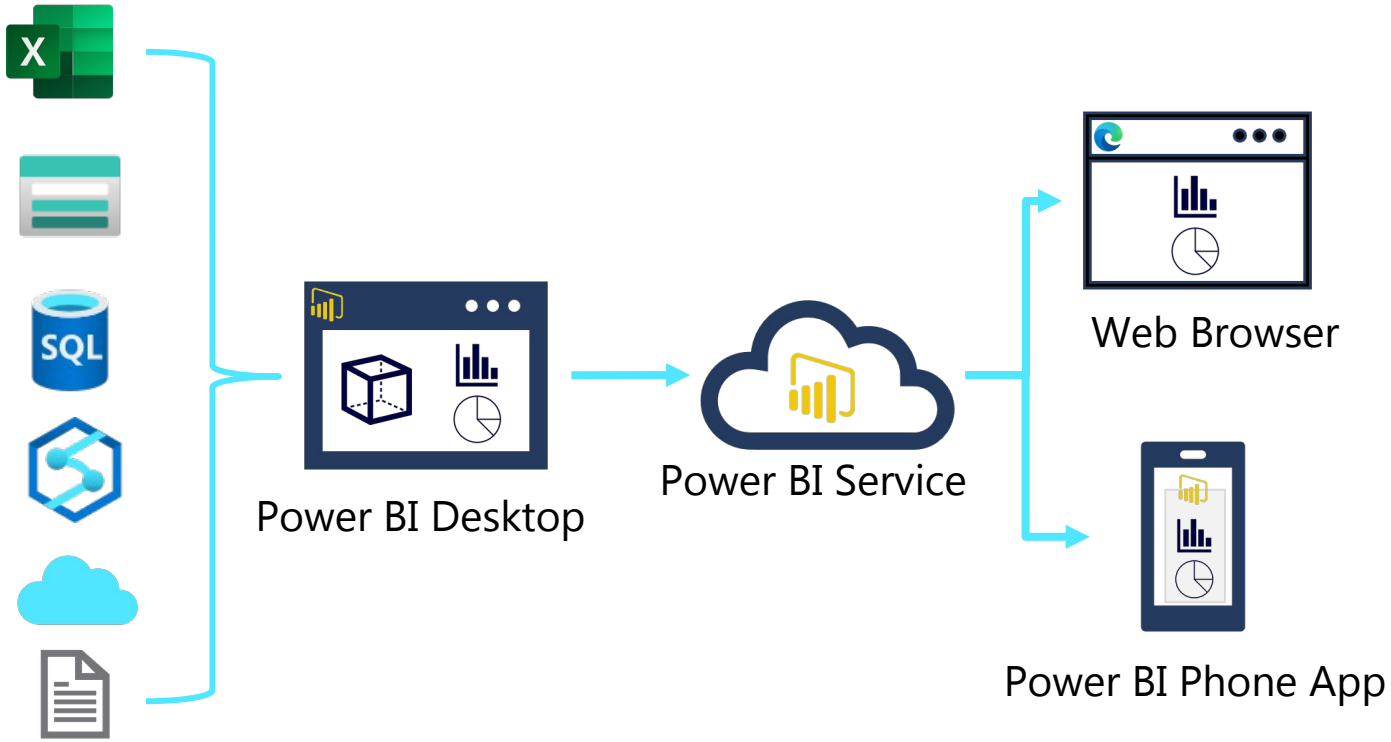
© Copyright Microsoft Corporation. All rights reserved.

3: Data visualization

© Copyright Microsoft Corporation. All rights reserved.

Introduction to data visualization with Power BI

- Start with Power BI Desktop
  - Import data from one or more sources
  - Define a data model
  - Create visualizations in a report
- Publish to Power BI Service
  - Schedule data refresh
  - Create dashboards and apps
  - Share with other users
- Interact with published reports
  - Web browser
  - Power BI phone app



© Copyright Microsoft Corporation. All rights reserved.

Analytical data modeling

Customer *(dimension)*

Key	Name	Address	City
1	Joe	1 Main St.	Seattle
2	Samir	123 Elm Pl.	New York
3	Alice	2 High St.	Seattle

Product *(dimension)*

Key	Name	Category
1	Hammer	Tools
2	Screwdriver	Tools
3	Wrench	Tools
4	Bolts	Hardware

Sales *(fact)*

Key	TimeKey	ProductKey	CustomerKey	Quantity	Revenue
1	01012022	1	1	1	2.99
2	01012022	2	1	2	6.98
3	02012022	1	2	2	5.98

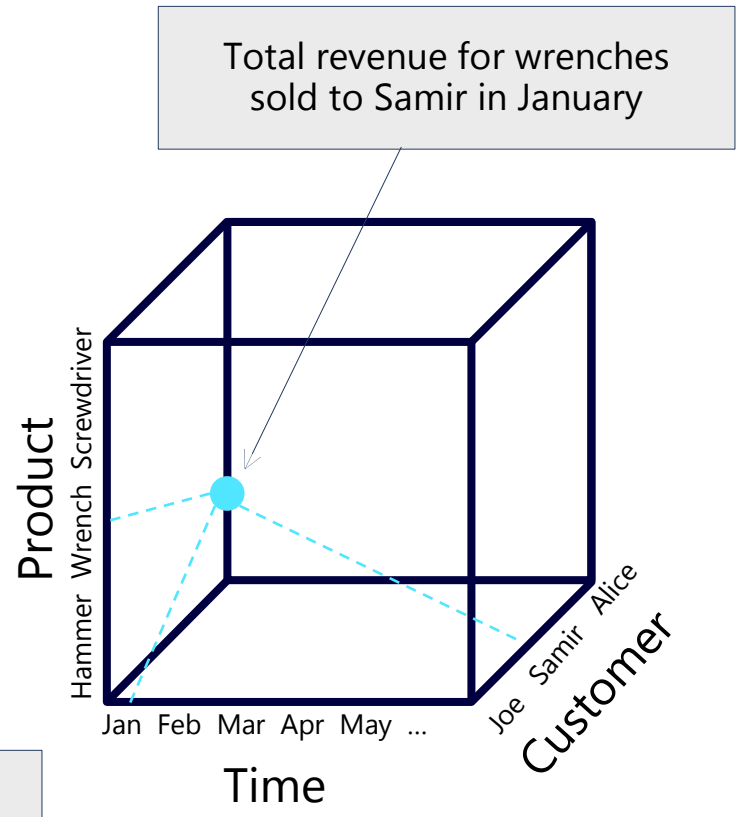
Measures

Time *(dimension)*

Key	Year	Month	Day	WeekDay
01012022	2022	Jan	1	Sat
02012022	2022	Jan	2	Sun

Hierarchy

∑



Model aggregates measures at each hierarchy level

Year	Month	Day	Revenue
2022			8221.48
	Jan		574.86
		1	9.97
		2	5.98
		...	...

© Copyright Microsoft Corporation. All rights reserved.

Common data visualizations in reports

Tables and text

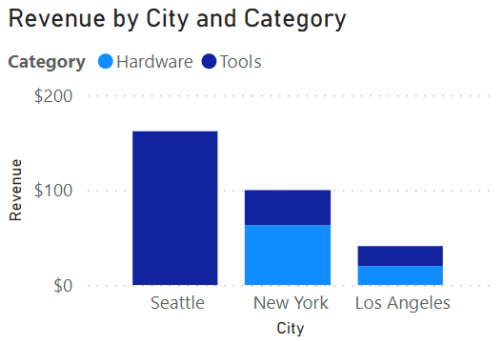
Product Sales

Name	Quantity
Bolts	2
Hammer	4
Nails	1
Screwdriver	2
Screws	2
Wrench	4
Total	**15**

$302.91

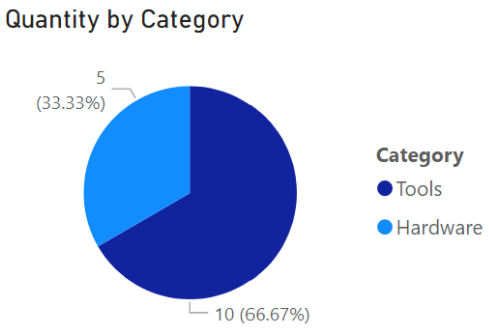
Revenue

Bar or column chart

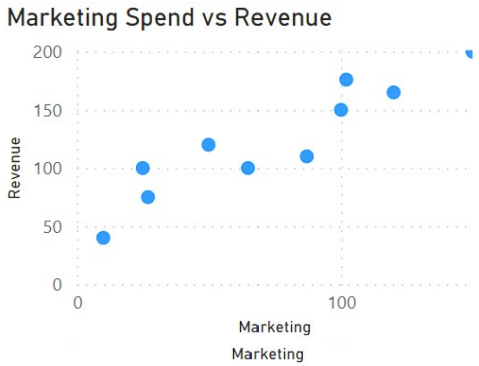


Line chart

Pie chart



Scatter plot



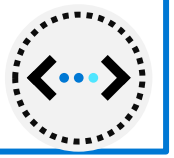
Map

© Copyright Microsoft Corporation. All rights reserved.

Lab: Visualize data with Power BI

In this lab, you will use Power BI Desktop to create a data model and a report

1. Start the virtual machine for this lab
   or go to the exercise page at https://aka.ms/dp900-pbi-lab
2. Follow the instructions to complete the exercise on Microsoft Learn
   Use the Azure subscription provided for this lab



© Copyright Microsoft Corporation. All rights reserved.

3: Knowledge check

Which tool should you use to import data from multiple data sources and create a report?

- [x] Power BI Desktop
- [ ] Power BI Phone app
- [ ] Azure Data Factory

What should you define in your data model to enable drill-up/down analysis?

- [ ] A measure
- [x] A hierarchy
- [ ] A relationship

Which kind of visualization should you use to analyze pass rates for multiple exams over time?

- [ ] A pie chart
- [ ] A scatter plot
- [x] A line chart

© Copyright Microsoft Corporation. All rights reserved.

Further learning

To review what you've learned and do additional labs, review the Microsoft Learn modules for this course:

- Explore core data concepts https://aka.ms/ExploreDataConcepts
- Explore relational data in Azure https://aka.ms/ExploreRelationalData
- Explore non-relational data in Azure https://aka.ms/ExploreNonRelationalData
- Explore data analytics in Azure https://aka.ms/ExploreDataAnalytics

© Copyright Microsoft Corporation. All rights reserved.

Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.